Performance of individual vs. group sampling for inferring dispersal under isolation-by-distance

Natacha Luximon, Eric Petit, Thomas Broquet

Performance of individual vs group sampling for inferring dispersal under isolation by distance

Natacha Luximon[1], Eric J. Petit[2], Thomas Broquet[3]

1- Department of Ecology and Evolution, University of Lausanne, CH-1015 Lausanne, Switzerland.

2- Univ Rennes 1, CNRS, UMR 6553 ECOBIO, Station Biologique, F-35380 Paimpont, France

3- Team Diversity and connectivity of coastal marine landscapes, UMR 7144, Station Biologique de Roscoff, CNRS & Univ Paris 06.

Keywords

Dispersal inference, gene flow, individual-based simulations, dispersal kernel, IBD

Corresponding author

Thomas Broquet, Team Div&Co, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France, phone number: +33 298 292 312, Email: thomas.broquet@sb-roscoff.fr

Running title

Group vs individual IBD analysis

24 Abstract

Models of isolation-by-distance formalize the effects of genetic drift and gene flow in a spatial
26 context where gene dispersal is spatially limited. These models have been used to show that, at an appropriate spatial scale, dispersal parameters can be inferred from the regression of genetic
28 differentiation against geographic distance between sampling locations. This approach is compelling because it is relatively simple and robust, and has rather low sampling requirements. In continuous
30 populations, dispersal can be inferred from isolation-by-distance patterns using either individuals or groups as sampling units. Intrigued by empirical findings where individual samples seemed to
32 provide more power, we used simulations to compare the performances of the two methods in a range of situations with different dispersal distributions. We found that sampling individuals
34 provides more power in a range of dispersal conditions that is narrow but fits many realistic situations. These situations were characterized not only by the general steepness of isolation-by-
36 distance but also by the intrinsic shape of the dispersal kernel. The performances of the two approaches are otherwise similar, suggesting that the choice of a sampling unit is globally less
38 important than other settings such as a study's spatial scale.

40 Introduction

Genetic data can inform us about dispersal patterns. But that information can be obtained only
42 when a number of biological and methodological conditions are fulfilled. At one end of a methodological continuum, the direct identification of dispersal events (e.g. using population or
44 parentage assignment) can provide detailed and accurate dispersal data. But because of its reliance on intensive sampling, this approach is constrained in terms of study systems, time frame, and study
46 area. At the other end, indirect estimates of migration rates obtained from measurements of spatial genetic structure and demogenetic models depend critically on models' refinement and assumptions

(Marko & Hart 2011; Whitlock & McCauley 1999). A sustained interest in this field of research has produced a wealth of alternative potential solutions for inferring dispersal (e.g. reviewed in Broquet & Petit 2009 and other references therein), but finding a good fit between biological settings and methodological options is rarely obvious.

Inferring dispersal from isolation-by-distance (IBD) patterns is one approach that seems to stand out by its (relatively) wide applicability. The dynamics of genetic variation in populations along a gradient of spatial proximity were first formalized by Wright (1943), Malécot (1949), and Kimura & Weiss (1964). These and following IBD theoretical developments have set ground for several inference methods that aim at estimating dispersal from genetic data (reviewed in Guillot *et al.* 2009). We focus here on the method proposed by Rousset (1997, 2000), which uses a regression of genetic distances on geographic distances among pairs of samples to infer the product $D\sigma^2$, where D is the effective density and σ^2 is the mean squared parent-offspring distance. If D can be independently estimated then σ^2 gives a synthetic descriptor of dispersal that can be compared across populations or species (e.g. Pinsky *et al.* 2010; see also Vekemans & Hardy 2004 using a related approach), and possibly compared with field-based estimates (e.g. Watts *et al.* 2007). The product $D\sigma^2$ itself is also of interest as it informs us on the increase of differentiation with distance. This approach is not free from drawbacks. Most importantly, the parameter σ is not intuitive (see discussions in Broquet & Petit 2009; Rousset 2004; Sumner *et al.* 2001), some preliminary knowledge of dispersal scale is needed to set an appropriate study scale, and data interpretation requires some understanding of the effect of departure from mutation-migration-drift equilibrium. But the method's robustness or behavior has been assessed in various aspects (e.g. Broquet *et al.* 2006b; Leblois *et al.* 2003; Leblois *et al.* 2004; Vekemans & Hardy 2004; Watts *et al.* 2007), and it relies on manageable sampling requirements. Accordingly, interpretations of isolation-by-distance patterns are frequent in the literature, including several estimations of the dispersal parameter σ (reviewed in Table S1, supplementary material. See also Fig. 1).

Rousset proposed to calculate distances between individuals in a continuous population (Rousset 2000) or between groups of individuals (either because the population under study is sub-divided into discrete units, or because discrete groups of individuals were sampled from an otherwise continuous population; Rousset 1997, 2000). Hereafter we will use the words "individual" and "group" to refer to the sampling unit of each approach. The two methods are based on the same theoretical background (detailed in Rousset 2004) and aim at estimating exactly the same quantity. Importantly, the two methods should be used at the same spatial scale, considering samples at distances not greater than ca. $0.56\sigma/\sqrt{2\mu}$, where μ is the mutation rate of the loci considered (Rousset 2004). Because the regression method based upon groups can be applied in a continuous population, some empirical case studies compared the results provided by the two methods with the same species in the same population (Broquet *et al.* 2006a; Suni & Gordon 2010; Watts *et al.* 2007). These studies repeatedly found that the group approach has less power, in some cases to the point that only the individual approach could be used to infer σ. The correlation of genetic and geographic distances is tested using Mantel's test, which is not particularly powerful (Legendre & Fortin 2010), and the number of pairwise comparisons is easily two orders of magnitude greater when using individuals as sampling units. The difference in power observed in case studies could thus be due simply to the number of data points, giving an advantage to individuals as sampling units. On the other hand, individual-based genetic distances may suffer from more sampling variance and more variable effect of genetic drift than group-based statistics. Differences in power remain to be investigated and complemented with results for the precision, bias, and coverage of confidence intervals obtained with each approach. The performances of IBD-based dispersal inference have been thoroughly evaluated in simulation studies that used individuals as sampling units (Leblois *et al.* 2003; Leblois *et al.* 2004). However, individual- and group-based sampling schemes have not yet been compared to one another in controlled conditions. Such a comparison could be useful for planning field studies and for interpreting empirical patterns, particularly in situations where samples are not easily collected individually. Such comparisons are also timely because of the

growing interest in using pooled samples (mixtures of individuals) that develops in parallel with
100 modern sequencing protocols (Davey *et al.* 2011; Futschik & Schlotterer 2010; Gautier *et al.* 2013).
Our objective is to determine whether there is an advantage in using one or the other method in
102 situations where the two methods could be applied.

104 Methods

Using IBDSim (Leblois *et al.* 2009) we simulated a continuous population composed by a square grid
106 of 110×110 units with one diploid individual per node. Each individual was characterized by a
multilocus genotype made of 10 microsatellites. IBDSim simulates the demography (coalescence and
108 dispersal) backwards in time before adding in mutations. The life-cycle is as follows: i) gamete
production and death of adults. ii) gamete mutation following a generalized stepwise model with
110 rate μ=5×10^{-4} as described in Leblois *et al.* (2004) with a maximum number of alleles set to 100 per
locus (a value large enough to be uninfluential here). iii) gamete dispersal according to a predefined
112 distribution of dispersal distances (see below). iv) constitution of diploid individuals. v) regulation of
the population to *n*=1 individual per node.

114 We defined 36 simulation scenarios (Table S1) differing only in dispersal conditions. Dispersal
distances followed a truncated Pareto distribution, where the probability of dispersing *k* steps in
116 each dimension is given by $f_k = M/k^n$ for $k \leq k_{max}$, as discussed by Rousset (2000). We varied *M*
(total dispersal rate in one dimension), *n* (a parameter that controls the shape of the distribution)
118 and k_{max} (maximum dispersal distance) to obtain a range of dispersal situations with simulated σ
values (σ^2_{sim}, range 1.12 – 47.01) comparable to that estimated from empirical case studies (σ^2_{est}: we
120 found estimates for 62 plant and animal species, Fig. 1 and Table S1). Simulations thus differed in the
values taken by σ^2 (giving the strength of IBD) but also in the nature of the dispersal kernels -
122 characterized by *M*, *n* and k_{max} - that yielded these values.

Dispersal inference under IBD should consider samples at distances smaller than ca. $0.56\sigma/\sqrt{2\mu}$ (Rousset 2004), which is approximately equal to 18σ given our mutation rate. The total size of the simulated population (110×110) was large enough to contain the optimal sampling design for any simulation scenario with some extra space to limit edge effects. At grid edges we used "absorbing" boundaries in IBDSim whereby "the probability mass of going outside the lattice is equally shared on all other movements inside the lattice" (as defined by R. Leblois in IBDSim user manual). The total simulated population was kept constant but samples were taken from within a smaller area and defined as a square of side length 13σ (that is, with diagonal $\approx 18\sigma$, Fig. 2). A different sampling grid was thus potentially associated with each simulation scenario.

To test for IBD and infer σ^2 we randomly sampled 99 individuals and 11 disjoint clusters of 9 individuals from within the defined sampling grid (Fig. 2). These samples were analyzed in Genepop V4.0 (Rousset 2008) using the estimator $\hat{a}$ for pairwise genetic distances among individuals (Rousset 2000) and $F_{ST}/(1-F_{ST})$ for groups (Rousset 1997). The mean genetic distance among pairs of samples and the global F_{ST} are shown in Table S2 and Fig. S1. The slope (b) of the regression of pairwise genetic distances and ln-transformed geographic distances among samples (individuals or groups) was used to infer σ^2 from the relationship $b = 1/(4D\pi\sigma^2)$ with D=1. We also recorded approximate 95% confidence intervals calculated using the ABC procedure implemented in Genepop (Leblois *et al.* 2003; Rousset 2008; Watts *et al.* 2007). Each simulation was replicated 200 times (a number large enough to capture most of the variance across replicates, data not shown), giving 36×200=7200 simulations overall.

The power of the regression method based upon groups and individuals was calculated for the 36 simulation conditions as the proportion of replicates yielding a significant Mantel test (using 10 000 permutations and a significance threshold α=5%). The relative error was estimated as $(\sigma^2_{est} - \sigma^2_{sim})/\sigma^2_{sim}$ for each replicate, and we defined the bias and the precision of σ^2 estimates as the median and the dispersion of the relative error, respectively. Finally, the coverage was defined

148 as the proportion of replicates where σ^2_{sim} was included within the confidence interval of σ^2_{est}. These statistics were computed using only the replicates where a significant IBD was detected (5045 and
150 5357 replicates for the group and individual methods), because σ^2 would not be inferred from a dataset otherwise.

152 We used generalized linear models to test for i) differences in power, bias, and coverage between methods, and ii) the effect of parameters *M*, *n*, and k_{max} on the power, bias, and coverage
154 of each method (with adequate transformation of data or binomial error structure when necessary). We included σ^2_{sim} as an explanatory variable in these models because it is directly linked to the
156 strength of IBD and thus should be a primary determinant of a method's performances. All results reported in the main text are thus independent of the value taken by σ^2_{sim}. To control for the fact
158 that different simulation conditions were associated with different sampling grids (i.e. different spatial scales), we also included the median of the Euclidean distances among pairs of samples as an
160 explanatory variable. Finally, the models were of the form < response ~ σ^2_{sim} + Med.dist + Method > when we compared the two methods (Med. dist is the median of distances among samples) and of
162 the form < response ~ σ^2_{sim} + Med. dist + $n \times M \times k_{max}$ > when we assessed the effects of simulation parameters, where response was either power, bias, or coverage.

164

Results

166 The power of the two methods, measured as the proportion of replicates yielding a significant Mantel test, dropped from 100% to ca. 20% in our two most extreme situations in terms of
168 simulated dispersal (Fig. 3a, σ^2_{sim}=1.12, and Fig. 3i, σ^2_{sim}=47.05). However, the group approach lost power at an earlier stage as the strength of IBD decreased (Figs. 3d-e). Interestingly, this effect was
170 primarily due to the shape parameter *n* (e.g. Figs. 3a,d,g), which had a significant effect independently of the value taken by σ^2_{sim} ($p<0.001$). When *n* was large (meaning that long-distance

dispersal was rare, first row in Fig. 3) the two methods performed well, and M and k_{max} took no effect. With low n the effect of k_{max} became critical (last row of Fig. 3) but the two methods were equally affected. When n was intermediate (middle row of Fig. 3) the group approach was more strongly affected than the individual approach by an increase in k_{max} (e.g. in Fig. 3d the power decreased from 100% to 80% for the individual approach vs 60% for the group approach when k_{max} was increased from 10 to 50). These results convey the following information: i) the two methods have comparable power except in a restricted set of conditions, ii) those dispersal conditions where individuals outperformed groups resulted in σ^2_{sim} in [3.89-11.83], a range of values that fits well empirical estimates from real case studies (Fig. 1), including one study where IBD was detected with individuals only (Broquet *et al.* 2006a), and iii) these conditions are not determined solely by σ^2 but also by the shape of the underlying dispersal kernel (e.g. the range of σ^2_{sim} mentioned above is also spanned by simulations 25-32, and yet with these simulations the two methods have nearly identical power, Fig. 3c&f).

Besides power, we looked at the bias and the precision of σ^2 estimates with the median and the dispersion of the relative error, respectively. We found that the two methods generally underestimated the true σ^2 by a small proportion (Fig. 4) and that this bias was slightly more pronounced with the individual approach (-15% and -9% for individuals and groups overall simulations, $p<0.001$). This slight underestimation is in agreement with simulation results obtained by Leblois *et al*. (2003; 2004) when the sampling design was not too far from theoretical optimum (e.g. simulations 1 and 2 in Table 2 of Leblois *et al*. 2003, note that the bias is calculated for the regression slope). In agreement with results for the power, the two methods showed decreasing performance (increasing bias) with decreasing IBD strength (down to ca. -60% with σ^2_{sim}=47.05, Fig. 4i). Irrespective of σ^2_{sim}, the bias also appeared to be influenced by the shape of the dispersal kernel, and particularly by parameter k_{max} ($p<0.001$). Increasing k_{max} resulted in deeper negative bias whatever the values taken by the other parameters. Surprisingly, the precision of estimates followed an opposite trend (Fig. 4): the dispersion of estimated values around the median was greatest when

198 IBD was strong, and this effect was particularly visible for small values of k_{max} (left box-plot of each panel in figure 4). As a result, the situations where the bias was minor were generally not favorable
200 in terms of precision. This observation is valid for the two methods, which showed no systematic difference in precision. Yet a difference can be noted regarding the replicates producing the worst
202 estimates. Overall simulations with significant IBD, 17 such replicates (out of 10 402) produced estimates with a relative error larger than 150% (Fig. 4). These cases were all characterized by a
204 near-zero slope estimate, yielding large relative errors. Interestingly, only 3 such cases were produced by the individual approach.

206 Finally, we did not find any difference in coverage between methods ($p>0.05$): the proportion of replicates where the 95% confidence interval of the estimate (σ^2_{est}) included the true value (σ^2_{sim})
208 amounted to 86% using groups and 85% using individuals (Fig. S2). In the specific cases where a difference in coverage was visible the method with the best coverage also appeared to have larger
210 confidence intervals (data not shown). Note that the coverage values reported here for each method independently may be overestimated, because the ABC procedure used to approximate 95%
212 confidence intervals generally underestimates the upper bound for σ^2_{est} (Leblois *et al.* 2003).

214 Discussion

Our simulations were parameterized so that the product $D\sigma^2$ fits real situations where IBD patterns
216 had been analyzed (Fig. 1 and Table S1). Yet the conditions of dispersal inference varied widely between simulations for the following reason: the number of samples was kept constant across
218 simulations (99 genotypes) while the sampling scale was set with respect to σ^2_{sim} in order to fit the methods' requirements (distance between samples $< 0.56\sigma_{sim}/\sqrt{2\mu}$). It means that the density of
220 the sampling effort decreased with increasing σ^2_{sim}, giving us a range of conditions where the inference of dispersal went from being very favored (when σ^2_{sim} is small and IBD is steep with

222 respect to the sampling scale) to very limited (with larger σ^2_{sim}). This variation allowed us to explore potential differences between the individual-based and group-based methods.

224 We find that there is only a small region of parameters where individual sampling outperformed group sampling, and this advantage bears upon power only (we found no sizeable
226 differences in accuracy, precision, and coverage between the two approaches). However, we note that intermediate situations, where the power of the individual-based regression approach was
228 greater than that of the group approach, appeared to cover the range of situations most commonly encountered in natural situations, at least in terms of $D\sigma^2$ (Fig. 1, exactly half of the reviewed
230 empirical estimates fall in the $D\sigma^2$region where the individual approach can outperform the group approach, depending on dispersal distributions).

232 Interestingly, the difference in performances between methods is due to particular conditions of σ^2_{sim} but also to the shape of the dispersal kernel (decreasing *n* significantly affected the
234 difference in power between methods independently of σ^2_{sim}, see Figs. 3a,d,g). Based upon empirical finding for a forest-dwelling mammal, the American marten, we had the intuition that dispersal
236 kernels characterized by a fat tail of long distance events could affect IBD patterns based upon groups more than individuals (Broquet *et al.* 2006a). But this idea is not supported by theory
238 (Rousset 2000), and our simulation results suggest that although there really is some effect of the shape of the dispersal kernel on the power of the two methods, it is not particularly due to long
240 distance dispersal.

We also found a slightly reduced risk to get extremely biased estimates with the individual approach
242 (considering those few estimates that were off by 150% or more, most came from group sampling). Furthermore, the accuracy of each method increased with the proportion of simulation replicates
244 where the two methods yielded a significant IBD pattern. This means that when one method yields a significant result but the other one does not then there is a higher risk of bias using either approach.
246 In other words, with adequate datasets that fulfill the methods' assumptions, the power difference

that may favor the individual-based approach occurs in situations where the risk of bias is anyways higher on average.

There are a number of relevant issues that were not considered here, such as the effects of spatial and temporal heterogeneity in population density on the relative performances of each approach (the density was set to 1 individual per node in all our simulations, see Leblois *et al.* 2003; Leblois *et al.* 2004 for different conditions with individual sampling). Whether or not such factors could interact with our findings is difficult to tackle, even using simplified simulations. Moreover, all our simulations fulfilled one critical assumption of IBD-based inferences (Rousset 1997, 2000): migration and drift are stable in space and time, and the pattern of increase of differentiation with geographic distance has reached equilibrium. The results presented here do not apply to other situations, which are irrelevant for inferring dispersal from IBD slopes, though the method seems robust to some disequilibrium situations (Leblois *et al.* 2004). Finally, we did not explore the effect of the number of samples (e.g. the number and the composition of groups). We chose to use rather small groups to get conservative results with the group approach, and because it is difficult to design simulation conditions that harmonize the requirements for sampling scale, useful σ_{sim}^2, and simulation and analysis time. In a pilot study we found nonetheless that increasing the total number of individuals sampled for each method benefited more to the group approach (data not shown).

Our findings suggest that when the methods are properly applied in continuously distributed populations there is only a slight advantage in using individuals as the sampling unit. Other considerations might thus be more important, such as the spatial scaling of IBD studies. As shown by previous work, the study scale should be large enough so that dispersal becomes spatially limited (unlike in the island model, which may apply at a shorter scale, e.g. see Kerth & Petit 2005), and, more critically, local enough so that the effect of gene flow does not faint out in front of mutation and is not blurred by non-equilibrated patterns (such as signatures of past colonization, e.g. Austin *et al.* 2004). Hence priority should be given to identifying the right study scale and choose the

sampling unit based upon the spatial distribution of individuals (Rousset 2000) and sampling possibilities rather than intrinsic properties of the methods. We emphasize that our conclusions about the detailed effect of dispersal parameters should not be extrapolated without caution to systems more complex than the simulations described here. But one robust result of this study is that in any case the choice of adequate spatial and temporal scales seems much more important than the sampling unit in continuously distributed populations.

Acknowledgements

We thank Raphaël Leblois for insightful discussions and comments, and for answering our questions regarding the software IBDSim. We are grateful to Glenn Yannic, Editor O. Gaggiotti and two anonymous reviewers for their constructive comments on the manuscript. We also thank Christophe Caron for his help with using the computer cluster in the biological station of Roscoff. TB was supported by the "Marine Aliens and Climate Change" program funded by AXA Research Funds.

References

Austin JD, Lougheed SC, Boag PT (2004) Controlling for the effects of history and nonequilibrium conditions in gene flow estimates in northern bullfrog (Rana catesbeiana) populations. *Genetics* **168**, 1491-1506.

Broquet T, Johnson CA, Petit E*, et al.* (2006a) Dispersal and genetic structure in the American marten, *Martes americana*. *Molecular Ecology* **15**, 1689-1697.

Broquet T, Petit E (2009) Molecular estimation of dispersal for ecology and population genetics. *Annual Review of Ecology, Evolution and Systematics* **40**, 193-216.

Broquet T, Ray N, Petit E, Fryxell JM, Burel F (2006b) Genetic isolation by distance and landscape connectivity in the American marten (*Martes americana*). *Landscape Ecology* **21**, 877-889.

296 Davey JW, Hohenlohe PA, Etter PD*, et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics* **12**, 499-510.

298 Futschik A, Schlotterer C (2010) The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics* **186**, 207-218.

300 Gautier M, Foucaud J, Gharbi K*, et al.* (2013) Estimation of population allele frequencies from next-generation sequencing data: pool- versus individual-based genotyping. *Molecular Ecology* 302 **22**, 3766-3779.

Guillot G, Leblois R, Coulon A, Frantz AC (2009) Statistical methods in spatial genetics. *Molecular* 304 *Ecology* **18**, 4734-4756.

Kerth G, Petit E (2005) Colonization and dispersal in a social species, the Bechstein's bat (*Myotis* 306 *bechsteinii*). *Molecular Ecology* **14**, 3943-3950.

Kimura M, Weiss GH (1964) The stepping stone model of population structure and the decrease of 308 genetic correlation with distance. *Genetics* **49**, 561-576.

Leblois R, Estoup A, Rousset F (2003) Influence of mutational and sampling factors on the estimation 310 of demographic parameters in a "continuous" population under isolation by distance. *Molecular Biology and Evolution* **20**, 491-502.

312 Leblois R, Estoup A, Rousset F (2009) IBDSim: a computer program to simulate genotypic data under isolation by distance. *Molecular Ecology Resources* **9**, 107-109.

314 Leblois R, Rousset F, Estoup A (2004) Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual 316 microsatellite data. *Genetics* **166**, 1081-1092.

Legendre P, Fortin MJ (2010) Comparison of the Mantel test and alternative approaches for 318 detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources* **10**, 831-844.

320 Malécot G (1949) Les processus stochastiques en génétique de population. *Publication de l'Institut de statistiques de l'Université de Paris I: Fasc* **3**, 1-16.

322 Marko PB, Hart MW (2011) The complex analytical landscape of gene flow inference. *Trends in Ecology & Evolution* **26**, 448-456.

324 Pinsky ML, Montes Jr. HR, Palumbi SR (2010) Using isolation by distance and effective density to estimated dispersal scales in anemonefish. *Evolution* **64**, 2688-2700.

326 Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219-1228.

328 Rousset F (2000) Genetic differentiation between individuals. *Journal of Evolutionary Biology* **13**, 58-62.

330 Rousset F (2004) *Genetic structure and selection in subdivided populations* Princeton University Press, Princeton.

332 Rousset F (2008) GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.

334 Sumner J, Rousset F, Estoup A, Moritz C (2001) Neighbourhood size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and

336 demographic methods. *Molecular Ecology* **10**, 1917-1927.

Suni SS, Gordon DM (2010) Fine-scale genetic structure and dispersal distance in the harvester ant

338 Pogonomyrmex barbatus. *Heredity* **104**, 168-173.

Vekemans X, Hardy O (2004) New insights from fine-scale genetic structure analyses in plant

340 populations. *Molecular Ecology* **13**, 921-935.

Watts PC, Rousset F, Saccheri IJ*, et al.* (2007) Compatible genetic and ecological estimates of

342 dispersal rates in insect (*Coenagrion mercuriale* : Odonata : Zygoptera) populations: analysis of 'neighbourhood size' using a more precise estimator. *Molecular Ecology* **16**, 737-751.

344 Whitlock MC, McCauley DE (1999) Indirect measures of gene flow and migration : $F_{ST\neq}1/(4Nm+1)$. *Heredity* **82**, 117-125.

346 Wright S (1943) Isolation by distance. *Genetics* **28**, 114-138.

348

Data accessibility

Three types of data are available on Dryad (entry doi:10.5061/dryad.mq8n5): i) parameter files for each simulation (IBDSim settings), ii) command lines for the batch analysis of simulation outputs in Genepop, and iii) summaries of the main statistics produced by Genepop for all simulations and replicates that we used here to compare the performances of individual-based vs group-based inferences.

Author contributions

N.L. performed the simulations and analyses. E.P. and T.B. designed the research, performed the analyses, and wrote the paper.

360 Figure legends

Figure 1 – Comparison of the frequency distribution of empirical (plotted as positive frequencies, in
362 grey) and simulated (represented by negative frequencies, in white) values of the product $D\sigma^2$. Empirical values were obtained from a literature survey of significant IBD patterns for animal and
364 plant case studies that investigated two-dimensional spatial genetic structure (the product $D\sigma^2$ has a different scale in 1D studies, Rousset 1997). These values were either taken directly from the papers,
366 or calculated from related statistics, such as Sp (Vekemans & Hardy 2004). When more than one value was available for a given species in a specific paper, only one was retained for drawing the
368 histogram. Most empirical values included in this comparison are taken from the review by Vekemans and Hardy (2004), completed with results from additional papers reviewed in our Table
370 S1 (supplementary material). Vertical lines show the lower and upper limits of the region in which individual-based analyses can outperform group-based analyses (see results). Note that the x axis is
372 log-scaled for a better visualization of the distributions.

374 Figure 2 – Principle of the sampling design. The actual simulations used a 110×110 grid, large enough to contain a square of side length 13σ for any of the conditions listed in Table S1. Ninety-nine
376 individuals or 11 groups of 9 individuals were randomly sampled from within this grid to infer σ^2 using isolation-by-distance patterns.

378

Figure 3 – Power of Mantel test in detecting a correlation between genetic and geographic distances
380 among pairs of individuals (dashed black lines) or groups (solid grey lines) sampled from simulated datasets. The power was calculated as the proportion of replicates (n=200 replicates per simulation
382 scenario) where a significant correlation was detected. Simulations differ in the distribution of dispersal distances (parameters M, n, and k_{max} of a truncated Pareto distribution).

384 Figure 4 –Relative error in σ^2 estimated from the regression of genetic- vs geographic distances between pairs of individuals (white boxes) or groups (grey). Data from 200 replicates per simulation
386 are shown (simulation conditions as in Figure 3). The solid line in each box shows the median of the error distribution, the box shows the 25% and 75% quantiles, and the whiskers show the full range of
388 the errors. In cases where the whiskers extend beyond the plotting region, 1 to 3 replicates (out of 200) had a relative error greater than 1.5 and are not shown here.

Fig. 1

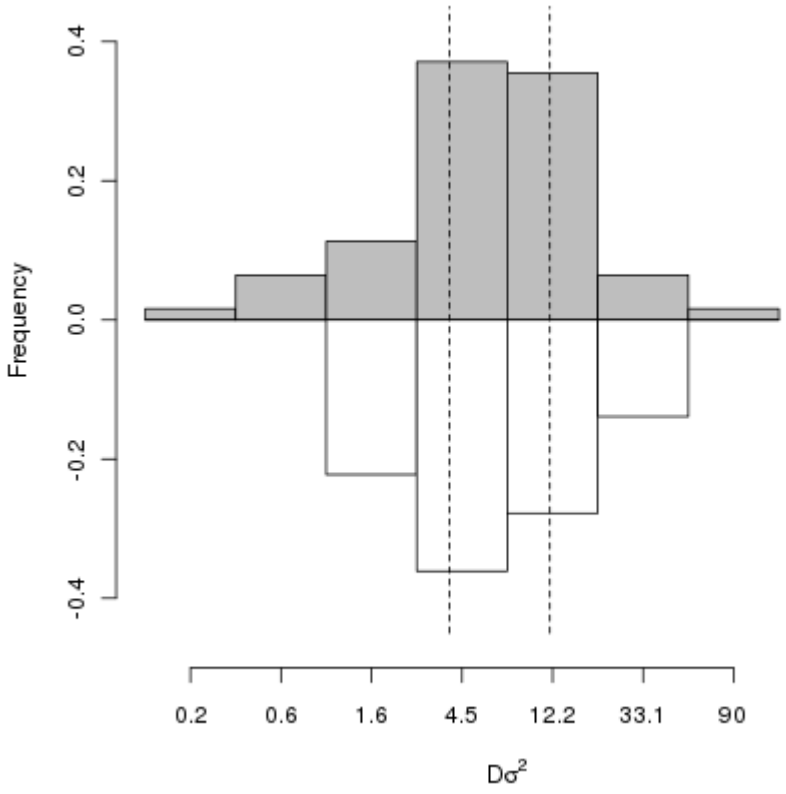
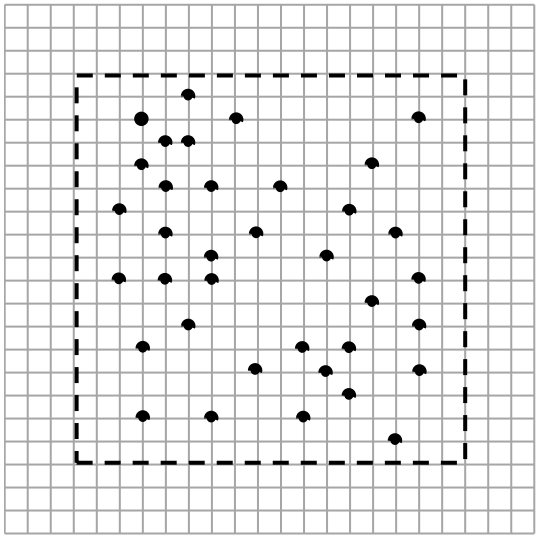


Fig. 2



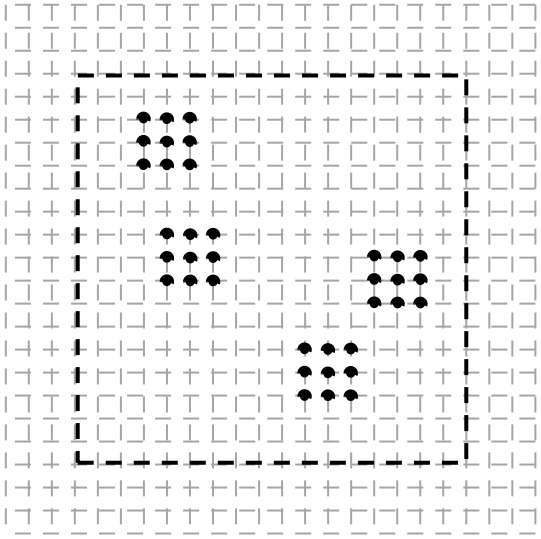


Fig. 3

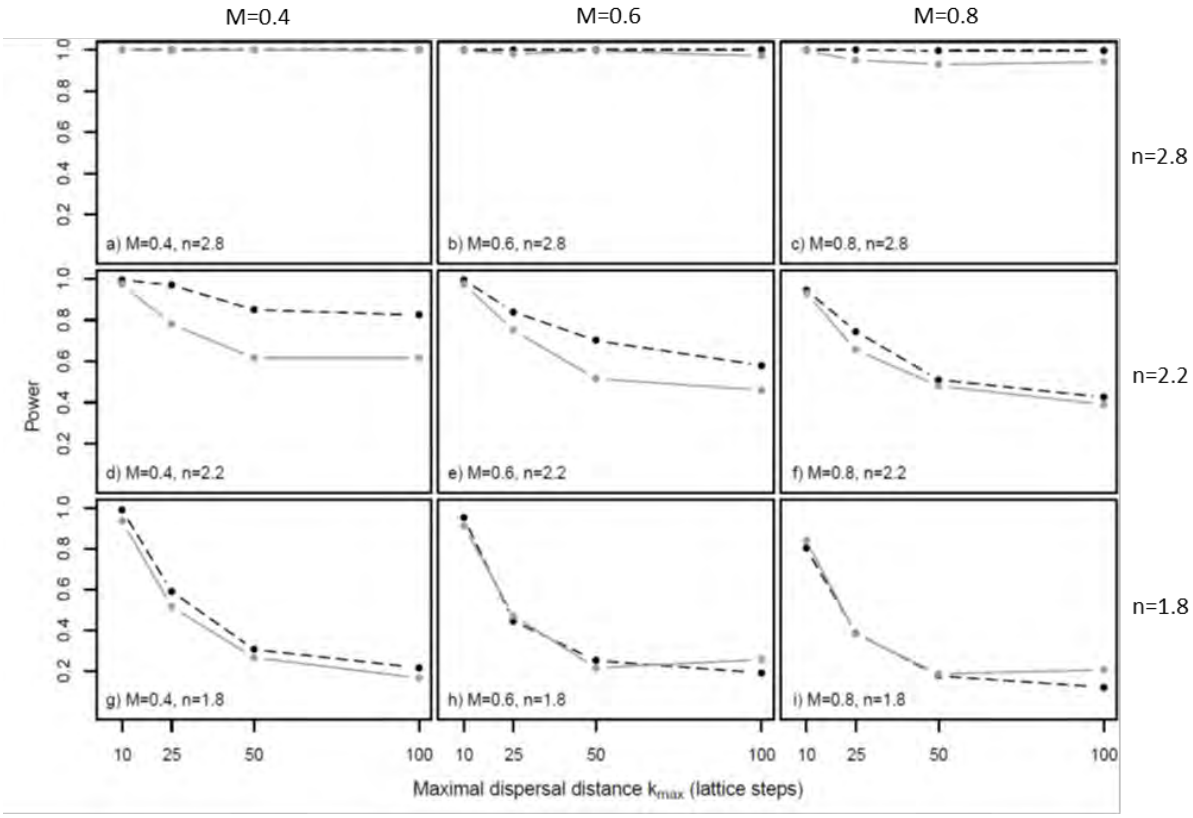
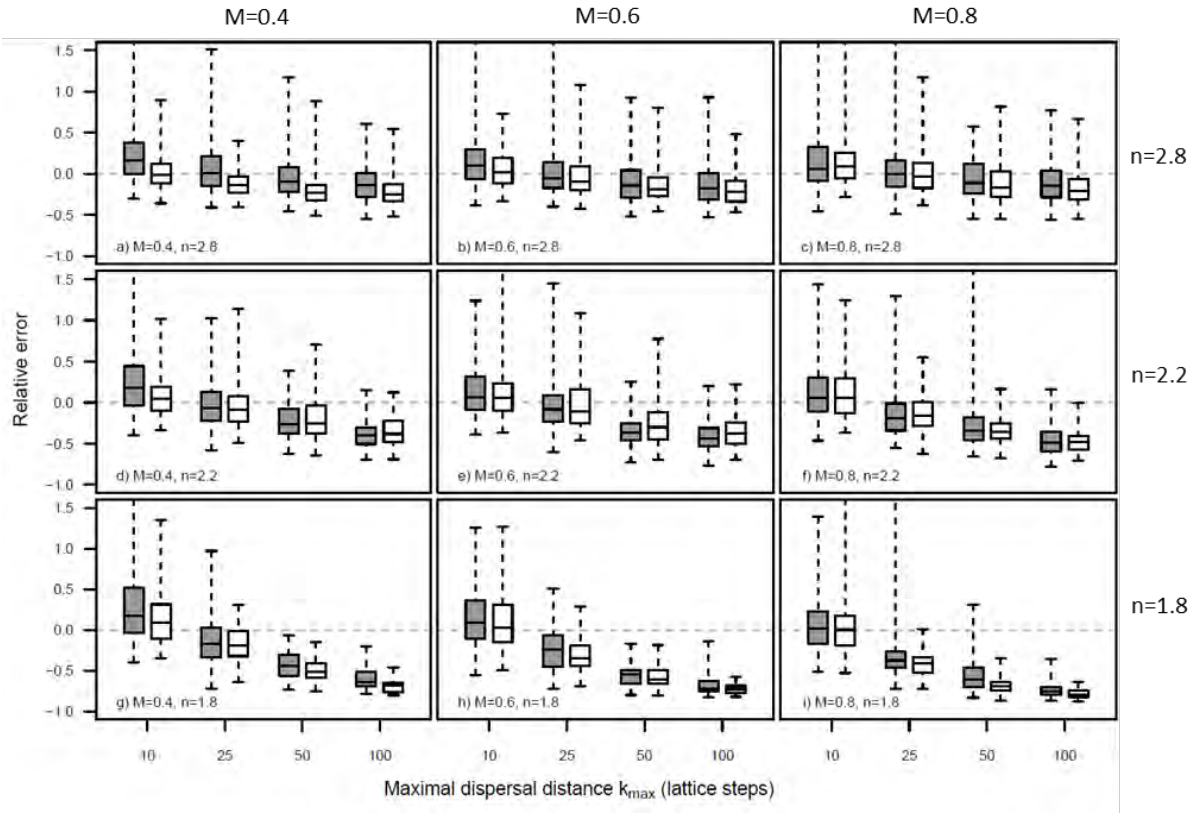

M=0.4
M=0.6
M=0.8
n=2.8
n=2.2
n=1.8
Power
a) M=0.4, n=2.8
b) M=0.6, n=2.8
c) M=0.8, n=2.8
d) M=0.4, n=2.2
e) M=0.6, n=2.2
f) M=0.8, n=2.2
g) M=0.4, n=1.8
h) M=0.6, n=1.8
i) M=0.8, n=1.8
0.2 0.4 0.6 0.8 1.0
10 25 50 100
Maximal dispersal distance k_{max} (lattice steps)

Fig. 4


M=0.4
M=0.6
M=0.8
n=2.8
n=2.2
n=1.8
Relative error
a) M=0.4, n=2.8
b) M=0.6, n=2.8
c) M=0.8, n=2.8
d) M=0.4, n=2.2
e) M=0.6, n=2.2
f) M=0.8, n=2.2
g) M=0.4, n=1.8
h) M=0.6, n=1.8
i) M=0.8, n=1.8
10
25
50
100
Maximal dispersal distance k_{max} (lattice steps)

Supplementary material

Table S1: Literature survey of empirical $D\sigma^2$ values estimated from significant IBD patterns for animal and plant case studies that investigated two-dimensional spatial genetic structure (the product $D\sigma^2$ has a different scale in 1D studies, Rousset 1997). These values were either taken directly from the papers, or calculated from related statistics, such as *Sp* (Vekemans& Hardy 2004). When more than one value was available for a given species in a specific paper, only one was retained for drawing the histogram (see footnotes).

Species	Taxon	Sampling unit	Density	Unit	$D\sigma^2$	Reference
Homo sapiens	Mammal	group	24	ind/km²	17	(Rousset 1997)
Dipodomys spectabilis	Mammal	individual	0.0002	ind/m²	2.6	(Rousset 2000)
Gnypetoscincus queenslandiae	Reptile	individual	0.0136	ind/m²	6.7	(Sumner *et al.* 2001)
Chamaecrista fasciculata	Fabaceae	group	-	-	10.7	(Fenster *et al.* 2003)
Ostrinia nubilalis	Insect	group	-	-	4.9	(Martel *et al.* 2003)
Crassostrea virginica	Mollusc	group	0.24	ind/km²	113.7	(Rose *et al.* 2006)
Martes americana	Mammal	individual	0.46	ind/km²	6.6	(Broquet *et al.* 2006)
Coenagrion mercuriale[a]	Insect	individual	0.0022	ind/m²	30.7	(Watts *et al.* 2007)
		group	0.0023	ind/m²	31.3	
Plethodon cinereus	Amphibian	group	2.82	ind/m²	9.9	(Cabe *et al.* 2007)
Microtus arvalis	Mammal	individual	1000	ind/km²	16.6	(Gauffre *et al.* 2008)
Bonasa bonasia	Bird	individual	5.5	ind/km²	5	(Sahlsten *et al.* 2008)
Milicia excelsia[b]	Moraceae (tree)	individual	4.96	ind/km²	12.3	(Bizoux *et al.* 2009)
Corrallium rubrum[c]	Cnidaria	group	-	-	8	(Ledoux *et al.* 2010)

(a) Based upon the data used to compare group- and individual-based approaches. (b) Computed from Table 2 for population Mindourou. (c) Computed from the slope value given in Fig. 2 for population Catalonia.

Table S2: Simulation conditions. The parameters *M*, *n*, and k_{max} set the shape of the distribution of dispersal distances in the simulations. The values taken by these parameters result in a range of dispersal conditions characterized by σ^2_{sim}. The genetic structure observed at equilibrium is given by the mean pairwise genetic distance between sampled individuals (*a*) or groups (F_{ST}/(1-F_{ST}) and the global F_{ST} averaged over 200 simulation replicates.

Simulation	*M*	*n*	K_{max}	σ^2_{sim}	mean genetic distance		
					a	F_{ST}/(1-F_{ST})	F_{ST}
1	0.4	2.8	10	1.12	0.202	0.103	0.093
2	0.4	2.8	25	1.53	0.201	0.101	0.091
3	0.4	2.8	50	1.83	0.203	0.103	0.093
4	0.4	2.8	100	2.01	0.200	0.103	0.093
5	0.4	2.2	10	1.99	0.159	0.071	0.066
6	0.4	2.2	25	3.89	0.151	0.064	0.060
7	0.4	2.2	50	6.03	0.151	0.063	0.059
8	0.4	2.2	100	7.85	0.154	0.064	0.060
9	0.4	1.8	10	3.08	0.133	0.050	0.047
10	0.4	1.8	25	7.93	0.120	0.040	0.038
11	0.4	1.8	50	15.23	0.115	0.037	0.036
12	0.4	1.8	100	23.13	0.119	0.037	0.035
13	0.6	2.8	10	1.68	0.101	0.074	0.068
14	0.6	2.8	25	2.30	0.101	0.074	0.068
15	0.6	2.8	50	2.75	0.104	0.075	0.069
16	0.6	2.8	100	3.02	0.104	0.075	0.069
17	0.6	2.2	10	2.99	0.071	0.051	0.048
18	0.6	2.2	25	5.85	0.065	0.043	0.041
19	0.6	2.2	50	9.08	0.063	0.042	0.040
20	0.6	2.2	100	11.83	0.064	0.042	0.040
21	0.6	1.8	10	4.63	0.050	0.035	0.034
22	0.6	1.8	25	11.95	0.041	0.026	0.025
23	0.6	1.8	50	22.99	0.038	0.023	0.022
24	0.6	1.8	100	34.98	0.041	0.023	0.023
25	0.8	2.8	10	2.24	0.077	0.057	0.054
26	0.8	2.8	25	3.08	0.072	0.052	0.049
27	0.8	2.8	50	3.68	0.075	0.053	0.050
28	0.8	2.8	100	4.03	0.076	0.055	0.052
29	0.8	2.2	10	4.00	0.047	0.036	0.034
30	0.8	2.2	25	7.83	0.040	0.030	0.029
31	0.8	2.2	50	12.16	0.039	0.028	0.027
32	0.8	2.2	100	15.84	0.039	0.029	0.028
33	0.8	1.8	10	6.20	0.031	0.024	0.023
34	0.8	1.8	25	16.02	0.020	0.016	0.016
35	0.8	1.8	50	30.86	0.018	0.015	0.015
36	0.8	1.8	100	47.05	0.020	0.015	0.014

Fig. S1 Mean pairwise genetic distance among individuals ($\hat{a}$, in black) or groups (F_{ST}/(1-F_{ST}), in grey) averaged over all replicates of each simulation. The results are classified according to simulation conditions: the panels differ in values taken by *M* and *n*, while the dots within each panel correspond to k_{max}= 10, 25, 50 and 100, respectively. The x-axis gives the resulting σ^2_{sim} averaged over all replicates of a simulation.

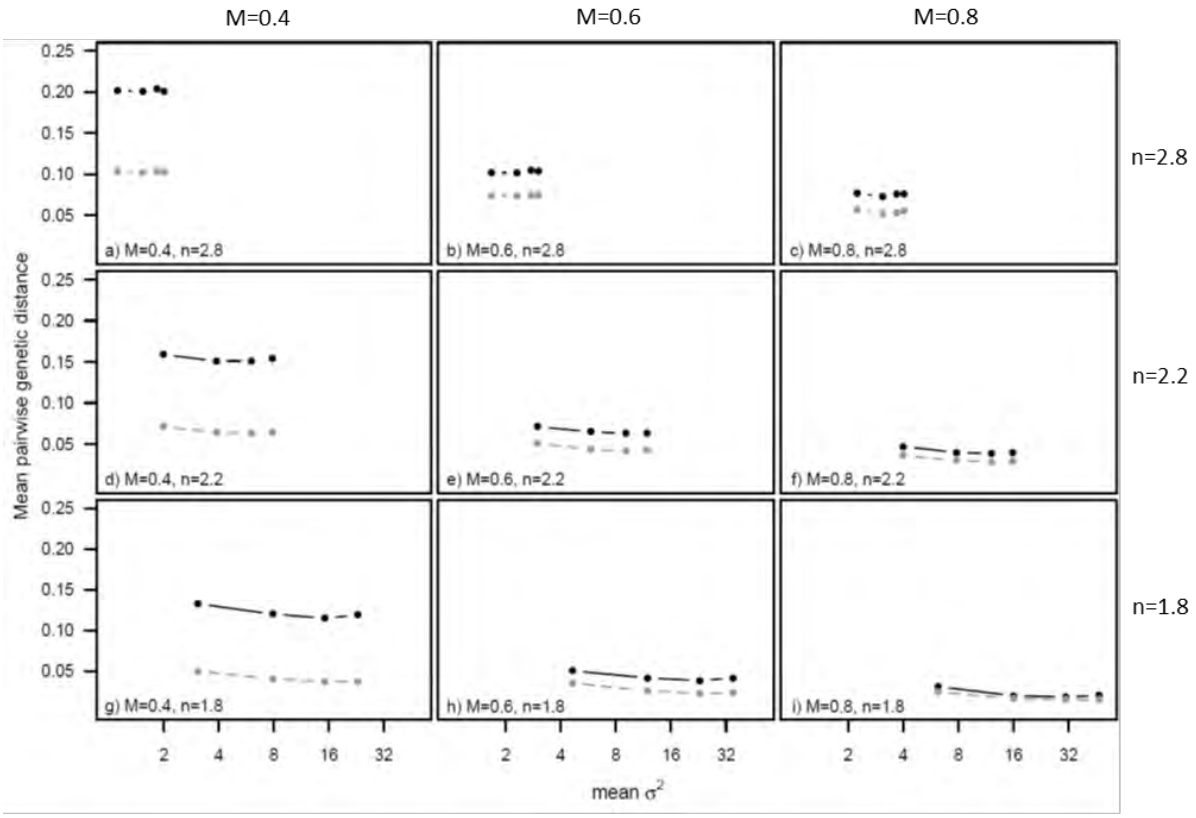
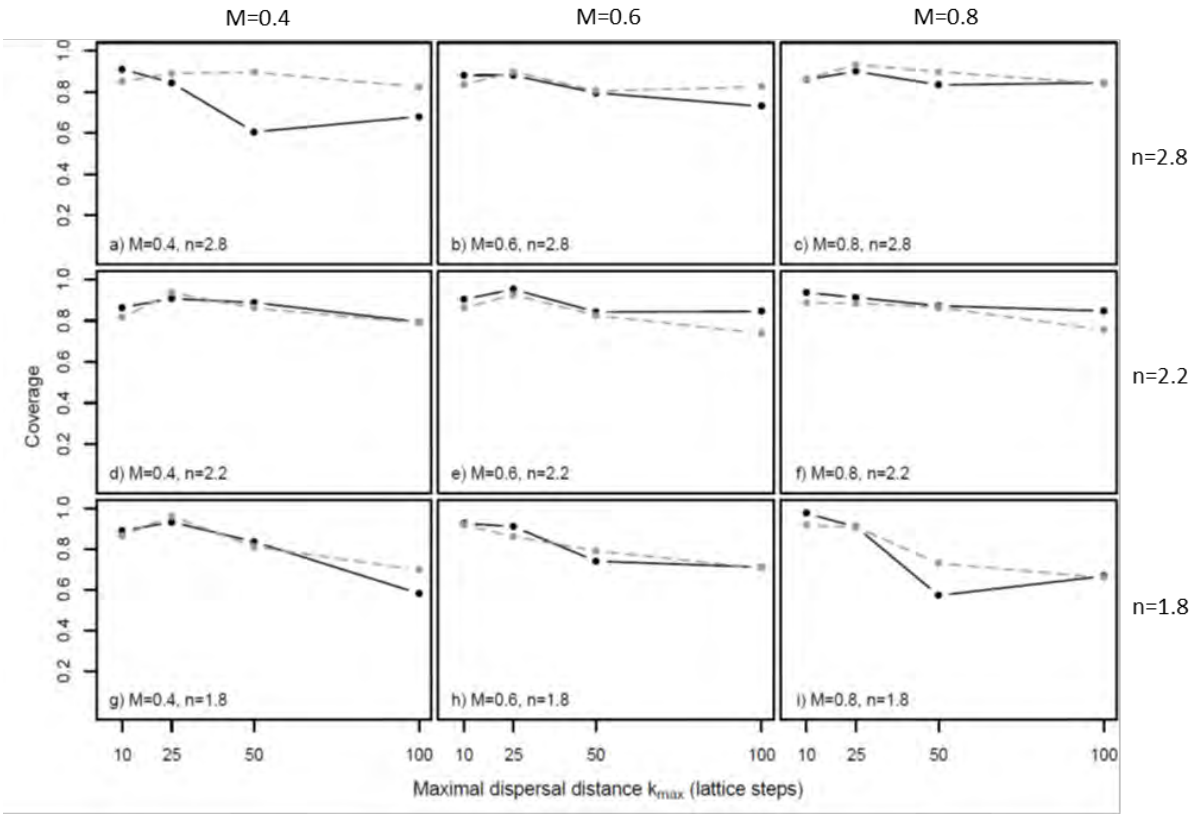


Fig. S2 –Coverage probability of 95% Confidence Intervals around σ^2 estimates from the regression of genetic- vs geographic distances between pairs of individuals (black) or groups (grey). Data from 200 replicates per simulation are shown (simulation conditions as in Figure S2).



References for supplementary material

Bizoux JP, Dainou K, Bourland N*, et al.* (2009) Spatial genetic structure in Milicia excelsa (Moraceae) indicates extensive gene dispersal in a low-density wind-pollinated tropical tree. *Molecular Ecology* **18**, 4398-4408.

Broquet T, Johnson CA, Petit E*, et al.* (2006) Dispersal and genetic structure in the American marten, *Martes americana*. *Molecular Ecology* **15**, 1689-1697.

Cabe PR, Page RB, Hanlon TJ*, et al.* (2007) Fine-scale population differentiation and gene flow in a terrestrial salamander (Plethodon cinereus) living in continuous habitat. *Heredity* **98**, 53-60.

Fenster CB, Vekemans X, Hardy O (2003) Quantifying gene flow from spatial genetic structure data in a metapopulation of *Chamaecrista fasciculata* (Leguminosae). *Evolution* **57**, 995-1007.

Gauffre B, Estoup A, Bretagnolle V, Cosson JF (2008) Spatial genetic structure of a small rodent in a heterogeneous landscape. *Molecular Ecology* **17**, 4619-4629.

Ledoux JB, Garrabou J, Bianchimani O*, et al.* (2010) Fine-scale genetic structure and inferences on population biology in the threatened Mediterranean red coral, Corallium rubrum. *Molecular Ecology* **19**, 4204-4216.

Martel C, Rejasse A, Rousset F, Bethenod MT, Bourguet D (2003) Host-plant-associated genetic differentiation in Northern French populations of the European corn borer. *Heredity* **90**, 141-149.

Rose CG, Paynter KT, Hare MP (2006) Isolation by distance in the eastern oyster, Crassostrea virginica, in Chesapeake Bay. *Journal of Heredity* **97**, 158-170.

Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* **145**, 1219-1228.

Rousset F (2000) Genetic differentiation between individuals. *Journal of Evolutionary Biology* **13**, 58-62.

Sahlsten J, Thorngren H, Hoglund J (2008) Inference of hazel grouse population structure using multilocus data: a landscape genetic approach. *Heredity* **101**, 475-482.

Sumner J, Rousset F, Estoup A, Moritz C (2001) Neighbourhood size, dispersal and density estimates in the prickly forest skink (*Gnypetoscincus queenslandiae*) using individual genetic and demographic methods. *Molecular Ecology* **10**, 1917-1927.

Watts PC, Rousset F, Saccheri IJ*, et al.* (2007) Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale* : Odonata : Zygoptera) populations: analysis of 'neighbourhood size' using a more precise estimator. *Molecular Ecology* **16**, 737-751.